

**GENERATING LARGE UNITS OF GRAPHEMES
WITH MUTUAL INFORMATION CRITERION FOR
LETTER TO SOUND CONVERSION**

5

ABSTRACT OF THE DISCLOSURE

A method and apparatus are provided for segmenting words into component parts. Under the invention, mutual information scores for pairs of grapheme units found in a set of words are
10 determined. Each grapheme unit includes at least one letter. The grapheme units of one pair of grapheme units are combined based on the mutual information score. This forms a new grapheme unit. Under one aspect of the invention, a syllable n-gram
15 model is trained based on words that have been segmented into syllables using mutual information. The syllable n-gram model is used to segment a phonetic representation of a new word into syllables. Similarly, an inventory of morphemes is formed using
20 mutual information and a morpheme n-gram is trained that can be used to segment a new word into a sequence of morphemes.